Milestone 2: Retrieval Component

Search Engine Implementation Report

This report presents the results of testing the search engine retrieval component with the required test queries. The search engine implements boolean AND queries with tf-idf scoring for ranking results.

Query Results

Query 1: cristina lopes

Found 5 results (Query time: 155.54 ms)

Rank	URL	Score
1	https://www.informatics.uci.edu/explore/faculty-profiles/cristina-lopes/	36.7490
2	https://www.informatics.uci.edu/2017/01/	27.1273
3	https://www.informatics.uci.edu/2017/10/	27.1273
4	https://www.informatics.uci.edu/2017/08/	27.1273
5	https://www.informatics.uci.edu/explore/facts-figures/	23.3190

Query 2: machine learning

Found 5 results (Query time: 0.44 ms)

Rank	URL	Score
1	https://www.cs.uci.edu/singh-to-deliver-machine-learning-talk-at-july-oc-acm-...	13.7771
2	https://www.cs.uci.edu/news/	12.8966
3	https://www.cs.uci.edu/professor-singh-awarded-two-nsf-grants-to-advance-mach...	12.8352
4	https://www.cs.uci.edu/news/page/16/	12.6115
5	https://www.cs.uci.edu/multidisciplinary-graduate-training-program-advances-w...	12.5501

Query 3: ACM

Found 5 results (Query time: 0.20 ms)

Rank	URL	Score
1	https://www.informatics.uci.edu/two-ics-professors-named-2015-acm-fellows/	12.6269
2	https://www.informatics.uci.edu/ics-alumni-named-to-acm-future-of-computing-a...	11.7370
3	https://www.informatics.uci.edu/update-franz-dourish-formally-recognized-as-a...	11.4636
4	https://www.informatics.uci.edu/informatics-professors-weave-sustainability-i...	10.8234
5	https://www.informatics.uci.edu/informatics-ph-d-student-saumya-gupta-receive...	10.4403

Query 4: master of software engineering

Found 5 results (Query time: 2.26 ms)

Rank	URL	Score
1	https://www.informatics.uci.edu/grad/mswe/	7.1250
2	https://www.informatics.uci.edu/grad/ms-software-engineering/	6.2402
3	https://www.informatics.uci.edu/grad/courses/	6.2002
4	https://www.informatics.uci.edu/undergrad/courses/	6.1264
5	https://www.informatics.uci.edu/training-tomorrows-software-engineers/	5.7955

Implementation Details

Query Processing

Queries are tokenized and stemmed using the same tokenizer and Porter stemmer used during indexing. This ensures query terms match indexed terms.

Boolean AND

The search engine implements boolean AND queries, meaning all query terms must be present in a document for it to be included in the results. The intersection of posting lists for all query terms is computed.

TF-IDF Scoring

Results are ranked using tf-idf (term frequency-inverse document frequency) scoring. The formula used is: tf_idf = (1 + log(tf)) * log(N/df), where tf is term frequency, df is document frequency, and N is the total number of documents.

Important Words

Words that appear in bold, headings (h1, h2, h3), or titles receive a 1.5x boost in their tf-idf score to reflect their higher importance.

Index Access

The search engine loads the inverted index into memory for fast query processing. Document mappings are also cached to enable fast URL lookups.